Deb8 - An ML and NLP-based Clickbait Detector

Romy Savin Peter
Computer Science & Engineering
IIIT Sri City
Andhra Pradesh, India

Emma Mary Cyriac
Computer Science & Engineering
IIIT Sri City
Andhra Pradesh, India

Riya Rajesh
Computer Science & Engineering
IIIT Sri City
Andhra Pradesh, India

Krushang Sirikonda
Computer Science & Engineering
IIIT Sri City
Andhra Pradesh, India

***Abstract* — This project integrates Natural Language Processing (NLP) and Machine Learning (ML) methodologies, including n-gram analysis, Term Frequency-Inverse Document Frequency (TF-IDF), tokenization, and stopword removal, to effectively categorize headlines as clickbait or non-clickbait. The goal is to demonstrate the feasibility of deploying this model on a larger scale, potentially serving as a robust tool in digital content management.**

1. INTRODUCTION

Clickbait, a term coined in the digital era, refers to web content with sensationalized headlines designed to attract reader attention and encourage clicks to a particular site. These headlines are often crafted without adherence to journalistic standards and are primarily motivated by generating advertising revenue and collecting user data. The prevalence of clickbait raises concerns about information quality and integrity in the digital space.

1.1 MOTIVATION

The digital age, characterized by the ubiquity of social media, smartphones, and high internet usage, has led to an overwhelming abundance of information. However, not all content holds equal value. The simplicity of content sharing and replication on social platforms has facilitated the proliferation of clickbait and misinformation. The economic incentives behind clickbait have further fueled its growth.

Addressing this issue, our project explores the potential of machine learning in identifying clickbait headlines. Our curiosity lies in understanding the intricacies of such a classification system. The primary objective is to develop a functioning clickbait detector, deployable as a web application, capable of discerning clickbait titles from genuine ones. This initiative is not merely an academic exercise but aims to demonstrate real-world applicability, potentially aiding in decluttering digital platforms by identifying and flagging dubious headlines, thereby enhancing the online experience for users.

2. CURRENT STATE OF THE WORK DONE IN THIS DOMAIN

There has been a great deal of research and work done in this domain and a few such popular papers that showcase the progress attained so far is mentioned below.

1. Olga Papadopoulou, Markos Zampoglou, Symeon Papadopoulos, & Ioannis Kompatsiaris. (2017). A Two-Level Classification Approach for Detecting Clickbait Posts using Text-Based Features.

   https://arxiv.org/abs/1710.08528

2. Naeem, B., Khan, A., Beg, M.O. *et al.* A deep learning framework for clickbait detection on social area network using natural language cues. *J Comput Soc Sc* **3,** 231–243 (2020).

   https://doi.org/10.1007/s42001-020-00063-y

3. A. Agrawal, "Clickbait detection using deep learning," 2016 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016, pp. 268-272

   https://ieeexplore.ieee.org/document/7877426

4. S. Chawda, A. Patil, A. Singh and A. Save, "A Novel Approach for Clickbait Detection," 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019, pp. 1318-1321

   https://ieeexplore.ieee.org/document/8862781

5. Ş. Genç and E. Surer, "Detecting “Clickbait” News on Social Media Using Machine Learning Algorithms," 2019 27th Signal Processing and Communications Applications Conference (SIU), 2019, pp. 1-4

   https://ieeexplore.ieee.org/document/8806257

6. A. Chakraborty, B. Paranjape, S. Kakarla and N. Ganguly, "Stop Clickbait: Detecting and preventing clickbaits in online news media," 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 2016, pp. 9-16

   https://ieeexplore.ieee.org/document/7752207

7. S. Manjesh, T. Kanakagiri, P. Vaishak, V. Chettiar and G. Shobha, "Clickbait Pattern Detection and Classification of News Headlines Using Natural Language Processing," 2017 2nd International Conference on Computational Systems and

Information Technology for Sustainable Solution (CSITSS), 2017, pp. 1-5

https://ieeexplore.ieee.org/document/8447715

8. J. Fu, L. Liang, X. Zhou and J. Zheng, "A Convolutional Neural Network for Clickbait Detection," 2017 4th International Conference on Information Science and Control Engineering (ICISCE), 2017, pp. 6-10

https://ieeexplore.ieee.org/document/8110177

9. M. Glenski, T. Weninger and S. Volkova, "Propagation from Deceptive News Sources Who Shares, How Much, How Evenly, and How Quickly?" in IEEE Transactions on Computational Social Systems, vol. 5, no. 4, pp. 1071-1082, Dec. 2018

https://ieeexplore.ieee.org/document/8542941

10. A. Bajaj, H. Nimesh, R. Sareen and D. K. Vishwakarma, "A Comparative Analysis of Classifiers Used for Detection of Clickbait in News Headlines," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1410-1415

https://ieeexplore.ieee.org/document/9432123

3. PROPOSED SYSTEM

The clickbait detection project is methodically designed to traverse through a sequence of distinct yet interconnected stages: data acquisition, data cleaning and processing, exploratory data analysis (EDA), and modeling & evaluation. Each phase plays a crucial role in shaping the efficacy of the final model.

3.1 DATA ACQUISITION

The project's cornerstone is an extensive dataset of approximately 52,000 headlines, meticulously compiled over the period from 2007 to 2020. This dataset integrates headlines from both clickbait and non-clickbait sources, ensuring a comprehensive and unbiased foundation for analysis.

➢ Sources:

- Kaggle Dataset (2007-2016): Incorporates 30,000 headlines, meticulously curated from the Clickbait Dataset by Aman Anand Rai.
- Scraped/Requested Headlines (2019-2020): Encompasses 22,000 headlines acquired through advanced scraping techniques and API requests, covering both social media platforms and online news publications.

➢ Diversity of Sources:

- Clickbait Sources: Headlines from renowned clickbait platforms like Thatscoop, Viralstories, Political Insider, Examiner, The Odyssey, Buzzfeed, Upworthy, Viral Nova, and Bored Panda.
- Non-Clickbait Sources: Credible news outlets such as The Guardian, Bloomberg, The Hindu, The New York Times, The Washington Post, Wiki News, and Reuters.

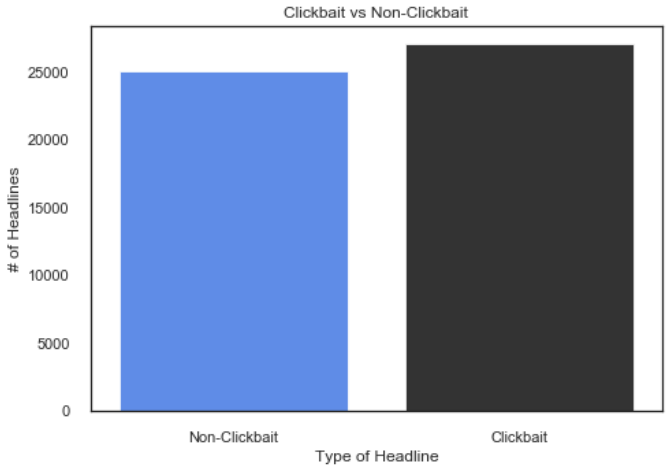


Fig.1 - Distribution of Clickbait vs. Non-Clickbait Headlines in the Dataset, visualized using Seaborn

3.2 CLEANING, PROCESSING & FEATURE ENGINEERING

In this phase, the headlines underwent a series of meticulous cleaning and processing steps:

- **Textual Cleanup:** Employing regular expressions, we removed punctuation, links, and non-alphabetical/non-numeric characters from headlines. Numbers were deliberately retained due to their frequent use in clickbait.
- **Stopword Removal:** A custom algorithm, leveraging a predefined list of frequent English stopwords, was employed to strip these non-essential words from the headlines.
- **Tokenization:** Initial Exploratory Data Analysis (EDA) involved tokenizing each headline, breaking down the text into individual words or tokens.

Beyond the text data, we engineered additional features to provide deeper insights:

1. **Headline Words:** Count of words in each headline, calculated prior to the removal of stopwords.
2. **Question Indicator:** Marks if a headline begins with a question word or contains a '?' (1 for 'yes', 0 for 'no'), computed before removing stopwords and punctuation.
3. **Exclamation Indicator:** Identifies the presence of an exclamation point (1 for 'yes', 0 for 'no'), determined before punctuation removal.
4. **Starts With Number:** Flags headlines starting with a number (1 for 'yes', 0 for 'no').

3.3 EXPLORATORY DATA ANALYSIS

Before progressing to modeling, we conducted an in-depth EDA to understand word frequencies, feature distributions, and the nuances between clickbait and non-clickbait headlines. This involved:

- **Vocabulary Analysis:** Assessing differences and overlaps in word usage between clickbait and non-clickbait content.
- **Feature Relevance:** Examining how engineered features impact classification.
- **Visualization:** Presenting findings through graphical representations for clearer insights.

3.4 MODELLING & EVALUATION

Multiple models were developed and evaluated to determine the most effective approach:

- **Model Selection:** Baseline dummy classifier, Naive Bayes, Random Forest, Linear SVM, and Logistic Regression models were tested.
- **NLP Techniques:** We integrated n-grams and TF-IDF for keyword extraction. Unigrams and bigrams were generated for all headlines, with corresponding TF-IDF scores aiding in the modeling process. Additional NLP steps included tokenization and stopword removal.
- **Evaluation Metrics:** Models were assessed based on accuracy and recall, with a preference for higher recall to reduce false negatives in clickbait identification.

4. RESULTS

Our analysis yielded intriguing insights, particularly in the lexical patterns distinguishing clickbait from non-clickbait headlines. Visual representations were key to our understanding and interpretation of these patterns.

4.1 RESULTS OF EXPLORATORY ANALYSIS

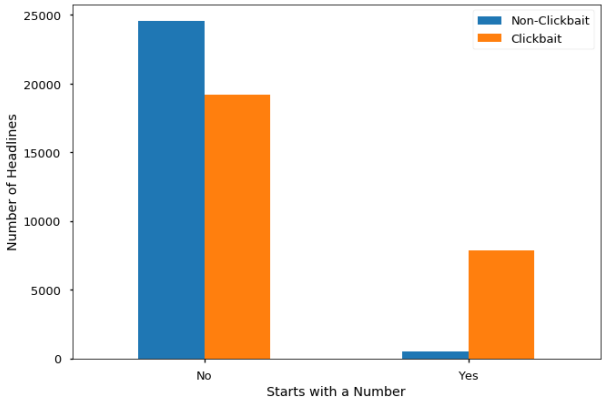
Fig.2: A WordCloud of the top 20 words in clickbait headlines, created using WordCloud and Matplotlib, reveals a distinct pattern of sensational and enticing language commonly used in clickbait.

Fig.3: Similarly, a WordCloud for non-clickbait headlines, illustrated through Matplotlib, showcases a more straightforward and informative linguistic style.

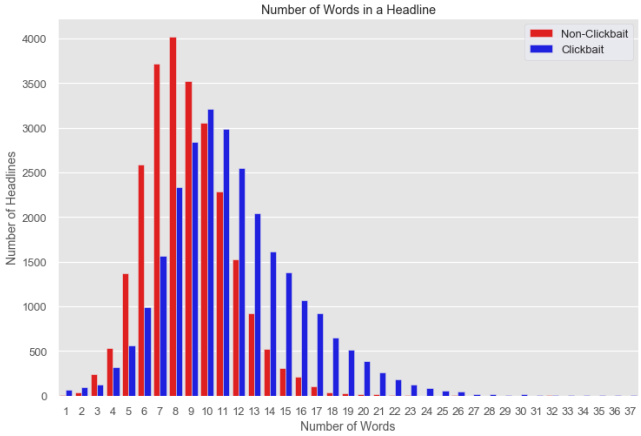
Our exploration into the engineered features presented notable differences between clickbait and non-clickbait headlines:

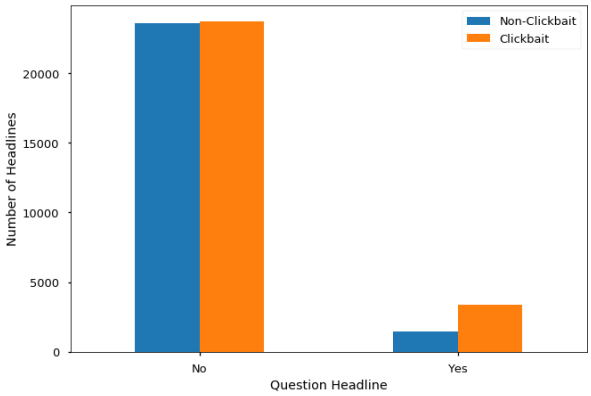
- **Headline Length:** On average, clickbait headlines tend to be slightly longer, possibly due to their elaborate and sensational nature.

- **Starting with a Number:** Clickbait headlines are more likely to start with a number, a common tactic to draw attention.
- **Question Presence:** The frequency of posing questions within headlines is higher in clickbait content, suggesting a strategy to provoke curiosity.



Shown below are the visualized results:





4.2 RESULTS OF MODELLING

In terms of modeling efficacy, the Naive Bayes classifier emerged as the most effective:

- **Performance Metrics:** As depicted in Fig.7, Naive Bayes outperformed other models in both recall and accuracy. However, the margins were narrow, indicating strong overall model performance across different algorithms.
- **Speed Advantage:** Naive Bayes' faster processing times make it a preferable choice in scenarios involving extensive data, aligning well with real-world applications.

	TEST ACCURACY	TEST RECALL
Baseline - Dummy Classifier	.52	1
Random Forest	.91	.93
Naive Bayes	.93	.94
Linear SVM	.93	.92
Logistic Regression	.93	.92

Fig.7

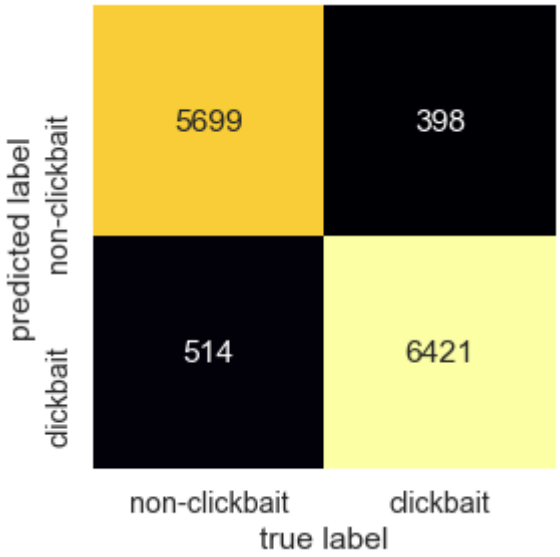
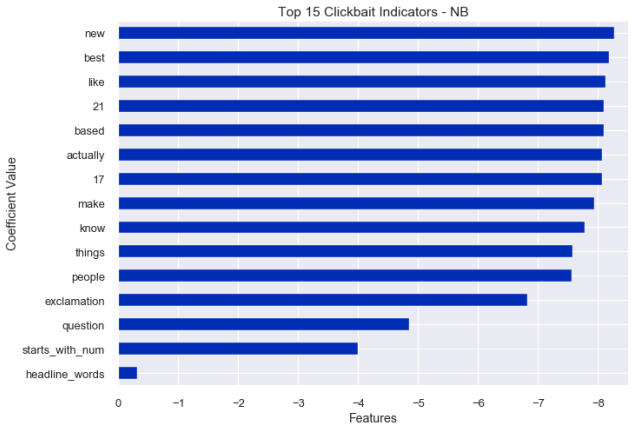


Fig.8

- **Naive Bayes Confusion Matrix:** Fig.8 illustrates the confusion matrix for the Naive Bayes model, providing a clear visual representation of its classification accuracy.
- **Coefficient Analysis:** Further insights were gained by examining the model's coefficients. This allowed us to understand which features and words the model prioritized in distinguishing between clickbait and non-clickbait headlines.



5. CONCLUSION & FUTURE WORK

5.1 CONCLUSION

Our project successfully leveraged a combination of Machine Learning algorithms and NLP techniques to distinguish between clickbait and non-clickbait headlines:

1. **Effective Classification:** Utilizing Naive Bayes, Logistic Regression, and SVM, coupled with NLP methodologies like tokenization, stopword removal, n-grams, and TF-IDF, we achieved accuracy and recall scores in the 90-93% range. A slight emphasis on recall was strategic, aimed at minimizing false negatives, i.e., wrongly classifying clickbait as non-clickbait.
2. **Real-World Application:** The success of these ML models in our study underlines the

feasibility of deploying such systems on a larger scale. This could revolutionize how digital platforms filter or flag clickbait content, offering preemptive solutions to readers.

3. **Model Insights:** In-depth analysis of the best-performing models' coefficients provided valuable insights into the classification mechanisms, enhancing our understanding of clickbait detection dynamics.
4. **Advancing Beyond Theory:** Building on existing theoretical frameworks, our implementation not only covers a diverse range of training data but also ensures consistency in feature sets, addressing common challenges in real-world scenarios. Additionally, our project goes a step further by integrating an aesthetically pleasing front-end UI, enhancing user experience and practicality.

5.2 FUTURE WORK

While our model represents a significant stride in clickbait detection, there is always room for growth and refinement. Potential avenues for future enhancements include:

- **Advanced NLP Techniques:** Delving into Deep NLP and Neural Network models could potentially yield a more robust classifier.
- **Thematic Analysis:** Employing Latent Dirichlet Allocation (LDA) to analyze topics and themes within headlines, which might offer new dimensions for classification.
- **LDA in Modeling:** Integrating LDA-derived topics into our modeling process could provide a richer feature set for classification.
- **Model Testing:** To further validate and refine our model, testing it on a new dataset would provide insights into its adaptability and scalability.

Our journey in developing an effective clickbait detection model demonstrates not only the potential of combining ML and NLP but also sets the stage for further innovations in this field.